Asymptotic and Constructive Bounds for Covering Arrays

Charles J. Colbourn[1]
with Erin Lanus and Kaushik Sarkar

[1]School of Computing, Informatics, and Decision Systems Engineering
Arizona State University

12 July 2017

Covering Array. Definition

- Let N, k, t, and v be positive integers.
- Let C be an $N \times k$ array with entries from an alphabet Σ of size v; we typically take $\Sigma = \{0, \ldots, v-1\}$.
- When $(\nu_1, \ldots, \nu_t)$ is a t-tuple with $\nu_i \in \Sigma$ for $1 \le i \le t$, $(c_1, \ldots, c_t)$ is a tuple of t column indices ($c_i \in \{1, \ldots, k\}$), and $c_i \neq c_j$ whenever $\nu_i \neq \nu_j$, the t-tuple $\{(c_i, \nu_i) : 1 \le i \le t\}$ is a *t-way interaction*.
- The array *covers* the t-way interaction $\{(c_i, \nu_i) : 1 \le i \le t\}$ if, in at least one row ρ of C, the entry in row ρ and column c_i is ν_i for $1 \le i \le t$.
- Array C is a *covering array* $\mathrm{CA}(N; t, k, v)$ of *strength t* when every t-way interaction is covered.
- $\mathrm{CAN}(t, k, v)$ is the minimum N for which a $\mathrm{CA}(N; t, k, v)$ exists.

Covering Array. Definition

- Let N, k, t, and v be positive integers.
- Let C be an $N \times k$ array with entries from an alphabet Σ of size v; we typically take $\Sigma = \{0, \ldots, v-1\}$.
- When $(\nu_1, \ldots, \nu_t)$ is a t-tuple with $\nu_i \in \Sigma$ for $1 \leq i \leq t$, $(c_1, \ldots, c_t)$ is a tuple of t column indices $(c_i \in \{1, \ldots, k\})$, and $c_i \neq c_j$ whenever $\nu_i \neq \nu_j$, the t-tuple $\{(c_i, \nu_i) : 1 \leq i \leq t\}$ is a *t-way interaction*.
- The array *covers* the t-way interaction $\{(c_i, \nu_i) : 1 \leq i \leq t\}$ if, in at least one row ρ of C, the entry in row ρ and column c_i is ν_i for $1 \leq i \leq t$.
- Array C is a *covering array* CA$(N; t, k, v)$ of *strength t* when every t-way interaction is covered.
- CAN(t, k, v) is the minimum N for which a CA$(N; t, k, v)$ exists.

Covering Array. Definition

- Let N, k, t, and v be positive integers.
- Let C be an $N \times k$ array with entries from an alphabet Σ of size v; we typically take $\Sigma = \{0, \ldots, v-1\}$.
- When $(\nu_1, \ldots, \nu_t)$ is a t-tuple with $\nu_i \in \Sigma$ for $1 \le i \le t$, $(c_1, \ldots, c_t)$ is a tuple of t column indices $(c_i \in \{1, \ldots, k\})$, and $c_i \ne c_j$ whenever $\nu_i \ne \nu_j$, the t-tuple $\{(c_i, \nu_i) : 1 \le i \le t\}$ is a *t-way interaction*.
- The array *covers* the t-way interaction $\{(c_i, \nu_i) : 1 \le i \le t\}$ if, in at least one row ρ of C, the entry in row ρ and column c_i is ν_i for $1 \le i \le t$.
- Array C is a *covering array* CA(N; t, k, v) of *strength t* when every t-way interaction is covered.
- CAN(t, k, v) is the minimum N for which a CA(N; t, k, v) exists.

Covering Array

CA(13;3,10,2)

0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
1	1	1	0	1	0	0	0	0	1
1	0	1	1	0	1	0	1	0	0
1	0	0	0	1	1	1	0	0	0
0	1	1	0	0	1	0	0	1	0
0	0	1	0	1	0	1	1	1	0
1	1	0	1	0	0	1	0	1	0
0	0	0	1	1	1	0	0	1	1
0	0	1	1	0	0	1	0	0	1
0	1	0	1	1	0	0	1	0	0
1	0	0	0	0	0	0	1	1	1
0	1	0	0	0	1	1	1	0	1

The Motivating Questions

1. How precisely can we determine $\mathrm{CAN}(t, k, v)$?
2. When we can show $\mathrm{CAN}(t, k, v) \leq N$, can we construct a $\mathrm{CA}(N; t, k, v)$ efficiently and explicitly?

A Random Method

- Fix t and v independent of k.
- In an array chosen uniformly at random from $\{0, \ldots, v-1\}^{N \times k}$, the probability that any specific t-way interaction is not covered is $\left(1 - \frac{1}{v^t}\right)^N$.
- So the expected number of uncovered t-way interactions is $\binom{k}{t} v^t \left(1 - \frac{1}{v^t}\right)^N$.
- When this expected number is less than 1, some array has all t-way interactions covered!

A Random Method

- Take logarithms of $\binom{k}{t} v^t \left(1 - \frac{1}{v^t}\right)^N < 1$ to get

$$\mathrm{CAN}(t, k, v) \leq \frac{t}{\log \frac{v^t}{v^t - 1}} \log k(1 + \mathrm{o}(1))$$

- ($\mathrm{CAN}(t, k, v) = \Omega(\log k)$ is easy: No two columns can be identical.)

Derandomizing

The Stein-Lovász-Johnson Method

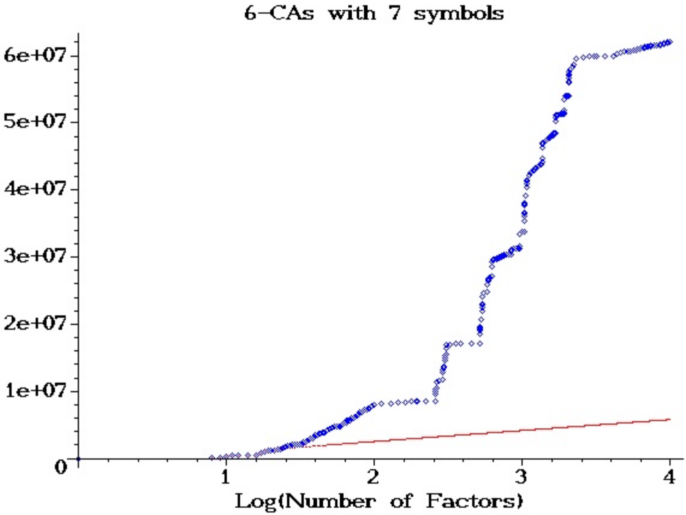
- Generate one row at a time at random from $\{0, \ldots, v-1\}^k$.
- The expected number of t-way interactions covered by this row for the first time is $\frac{1}{v^t}$ times the number of as-yet-uncovered t-way interactions.
- Stein (1974), Lovász (1975), and Johnson (1974): Select a row that covers the maximum number of as-yet-uncovered t-way interactions.
- But finding such a row is NP-hard!
- So select a row that covers at least the average.
- In fact, we do better: After each row is selected, the number of uncovered interactions is an integer. (Discrete SLJ)

Derandomizing

The Stein-Lovász-Johnson Method

- Generate one row at a time at random from $\{0, \ldots, v-1\}^k$.
- The expected number of t-way interactions covered by this row for the first time is $\frac{1}{v^t}$ times the number of as-yet-uncovered t-way interactions.
- Stein (1974), Lovász (1975), and Johnson (1974): Select a row that covers the maximum number of as-yet-uncovered t-way interactions.
- But finding such a row is NP-hard!
- So select a row that covers at least the average.
- In fact, we do better: After each row is selected, the number of uncovered interactions is an integer. (Discrete SLJ)

Computational Results



Better asymptotics

LLL

- SLJ and Discrete SLJ do not account for the limited statistical dependence among the events of coverage of interactions.
- The (symmetric version of the) Lovász Local Lemma (LLL) yields a better bound (obtained by Godbole, Skipper, and Sunley in 1996)

$$\mathrm{CAN}(t,k,v) \leq \frac{t-1}{\log \frac{v^t}{v^t-1}} \log k(1+\mathrm{o}(1))$$

Better asymptotics

Francetic-Stevens

- Francetic and Stevens (2016) made the first improvement in 20 years, using an entropy compression technique

$$\mathrm{CAN}(t,k,v) \leq \frac{v(t-1)}{\log\left(\frac{v^{t-1}}{v^{t-1}-1}\right)} \log k(1+\mathrm{o}(1))$$

- Is it better? Use the Taylor series expansion to verify.

Constructive algorithms

- Applications require explicit constructions of arrays, not asymptotic bounds.
- Can we meet the bounds efficiently when t and v are fixed?
  - Discrete SLJ: Yes, an efficient conditional expectation method ("density") deterministically chooses a row as good as average (Bryce-C, 2007, 2009)
  - LLL: Yes if you allow expected polynomial time: Moser-Tardos (2010) give a resampling method that succeeds within a linear expected number of resamplings.
  - Francetic-Stevens: Not clear (yet), but stay tuned.

Constructive algorithms

Why are the tables so bad?

- When $v = 7$, $t = 6$, and $k = 50$ there are

  $$18695249643000$$

  interactions to cover!
- Density stores coverage information for each, and the storage requirement is enormous.
- Moser-Tardos recomputes coverage for each for every resampling, and the number of resamplings needed is a random variable.

Constructive algorithms

Sample space reduction

- Consider covering arrays that are invariant under the action of a group on the *symbols* of the array, in order to make the space to search for an array much smaller.
- We consider three permutation groups acting on the symbols.
  - the cyclic group of order v, which partitions the interactions on t columns into v^{t-1} orbits of length v;
  - the Frobenius or affine group when v is a prime power, which partitions the interactions into $\frac{v^{t-1}-1}{v-1}$ orbits of length $v(v-1)$ and one orbit of length v;
  - PGL when $v+1$ is a prime power, which partitions the interactions into orbits of length $v(v-1)(v-2)$, $v(v-1)$, and v.

Constructive algorithms

Covering Orbits

- Now we cover orbits of interactions and apply the group to recover the covering array at the end.
- We can apply the SLJ paradigm and the density methods in the same way in the cyclic and Frobenius cases (For density, see Colbourn 2013).
- We can apply LLL and the Moser-Tardos methods in the same way in the cyclic and Frobenius cases.
- This reduces time and storage for density, and time for Moser-Tardos — But what does it do to the asymptotic bounds?

Better asymptotics

Cyclic LLL

- Applying LLL with the cyclic group, we reproduce the Francetic and Stevens (2016) bound

$$\mathrm{CAN}(t,k,v) \leq \frac{v(t-1)}{\log\left(\frac{v^{t-1}}{v^{t-1}-1}\right)} \log k(1+\mathrm{o}(1))$$

- and we get a Moser-Tardos type method that runs in expected polynomial time to meet the bound.

Better asymptotics
Frobenius LLL

- Applying LLL with the Frobenius group, we improve on the Francetic and Stevens (2016) bound

$$\mathrm{CAN}(t,k,v) \leq \frac{v(v-1)(t-1)}{\log\left(\frac{v^{t-1}}{v^{t-1}-v+1}\right)} \log k(1+\mathrm{o}(1))$$

- and we get a Moser-Tardos type method that runs in expected polynomial time to meet the bound.

What about PGL?

- Covering orbits of length v can be done with v constant rows.
- Covering orbits of length $v(v-1)(v-2)$ can be done with LLL (or Moser-Tardos).
- But orbits of length $v(v-1)$ are a problem, in that their probability of being covered in a random selection is much smaller.
- So the road to higher levels of sharply ℓ-transitive groups acting on the symbols seems blocked.

Covering Perfect Hash Families Setup I

- George Sherwood suggested a framework for constructing covering arrays using finite fields.
  - q a prime power,
  - $\mathbb{F}_q$ the finite field of order q,
  - $\mathcal{R}_{t,q} = \{\mathbf{r}_0, \ldots, \mathbf{r}_{q^t-1}\}$ the set of all (row) vectors of length t with entries from $\mathbb{F}_q$
  - $\mathcal{T}_{t,q}$ the set of all nonzero column vectors of length t with entries from $\mathbb{F}_q$.
- A vector $\mathbf{x} \in \mathcal{T}_{t,q}$ is sometimes called a *permutation vector*.

Covering Perfect Hash Families Setup II

Lemma

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ be a set of vectors from $\mathcal{T}_{t,q}$. The array $A = (a_{ij})$ formed by setting a_{ij} to be the product of $\mathbf{r}_i$ and $\mathbf{x}_j$ is a $\mathrm{CA}(q^t; t, t, q)$ *if and only if the $t \times t$ matrix $X = [\mathbf{x}_1 \cdots \mathbf{x}_t]$ is nonsingular.*

- This is essentially the Bose-Bush construction of orthogonal arrays.

CPHFs and Covering Arrays

- A *covering perfect hash family* $\mathrm{CPHF}(n; k, q, t)$ is an $n \times k$ array $C = (\mathbf{c}_{ij})$ with entries from $\mathcal{T}_{t,q}$ so that, for every set $\{\gamma_1, \ldots, \gamma_t\}$ of distinct column indices, there is at least one row index ρ of C for which $[\mathbf{c}_{\rho\gamma_1} \cdots \mathbf{c}_{\rho\gamma_1}]$ is nonsingular; call this a *covering t-set* and say that the t-set of columns is *covered* in row ρ.

Lemma

Suppose that C is a $\mathrm{CPHF}(n; k, q, t)$. Then there exists a $\mathrm{CA}(n(q^t - 1) + 1; t, k, q)$.

CPHFs and Covering Arrays

- A *covering perfect hash family* CPHF$(n; k, q, t)$ is an $n \times k$ array $C = (\mathbf{c}_{ij})$ with entries from $\mathcal{T}_{t,q}$ so that, for every set $\{\gamma_1, \ldots, \gamma_t\}$ of distinct column indices, there is at least one row index ρ of C for which $[\mathbf{c}_{\rho\gamma_1} \cdots \mathbf{c}_{\rho\gamma_1}]$ is nonsingular; call this a *covering t-set* and say that the t-set of columns is *covered* in row ρ.

Lemma

Suppose that C is a CPHF$(n; k, q, t)$. *Then there exists a* CA$(n(q^t - 1) + 1; t, k, q)$.

CPHF Better Asymptotics for Covering Arrays

- Choose entries of an $n \times k$ array A uniformly at random from $\mathcal{T}_{t,q}$.
- Let T be a set of t columns of A.
- Within one row of A, the probability that the columns of T are *not* covering is

$$\phi_{t,q} := 1 - \frac{\prod_{i=0}^{t-1}(q^t - q^i)}{(q^t-1)^t} = 1 - \prod_{i=1}^{t-1} \frac{q^t - q^i}{q^t - 1}.$$

- The probability that A does not contain a covering t-set for T is $\phi_{t,q}^n$.

Random Selection with Postprocessing

- Constructing an array with $\kappa > k$ columns, we can compute the expected number of t-tuples of columns not covered.
- Choose κ so that the number of uncovered tuples of columns is $\kappa - k$.
- Then delete one column from each uncovered tuple.
- At least k remain, and the result is a covering array!

CPHF Better Asymptotics for Covering Arrays

Lemma

For all $q \geq 3$ and $t \geq 3$,

$$\frac{1}{q} \leq \phi_{t,q} \leq \frac{q+1}{q^2}.$$

- This leads to three better asymptotic bounds for covering arrays!

SLJ CPHF	LLL CPHF	SLJ (pp) CPHF
$\frac{v^t t}{2\log v - \log(v+1)}$	$\frac{v^t(t-1)}{2\log v - \log(v+1)}$	$\frac{v^t(t-1)}{2\log v - \log(v+1)}$

Next Steps: The Binary Case

- The CPHF approach improves known bounds when $v > 2$, but what about the binary case?
- Here Hadamard matrices lead to many of the best known bounds.
- One can view permutation vectors as (specific) functions of t variables over $\mathbb{F}_q$. The key is that we can determine the probability with which t such functions form a covering set.
- Can one find a set of t binary functions of s variables yielding improvements on the known bounds?
- It appears to me that Hadamard-type approaches are the appropriate methods here.